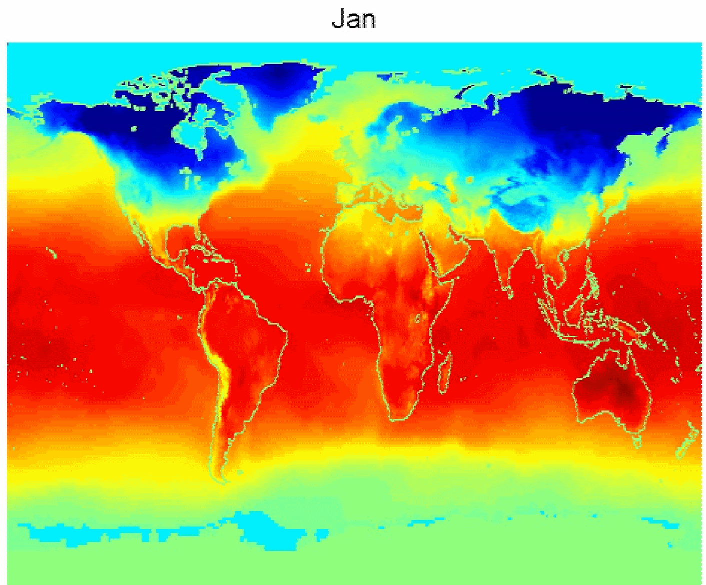
Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining

Principal Investigator: Vipin Kumar, University of Minnesota

Co-Investigators:

George Karypis	University of Minnesota
Steven Klooster	NASA Ames Research Center
Shashi Shekhar	University of Minnesota
Christopher Potter	NASA Ames Research Center

Discovery of Patterns in the Earth Science Data

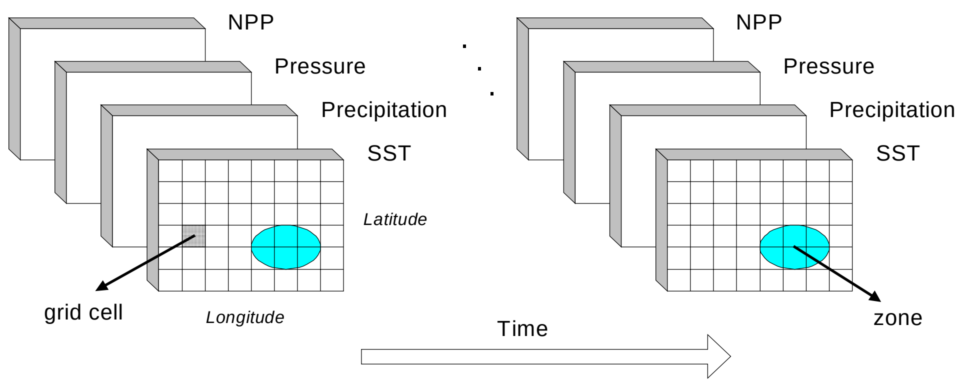


Goal: Better understand global scale patterns in biosphere processes, especially relationships between the global carbon cycle and the climate system.

Objectives: Develop data mining techniques to efficiently find spatio-temporal patterns in large Earth Science data sets.

Key Innovations:

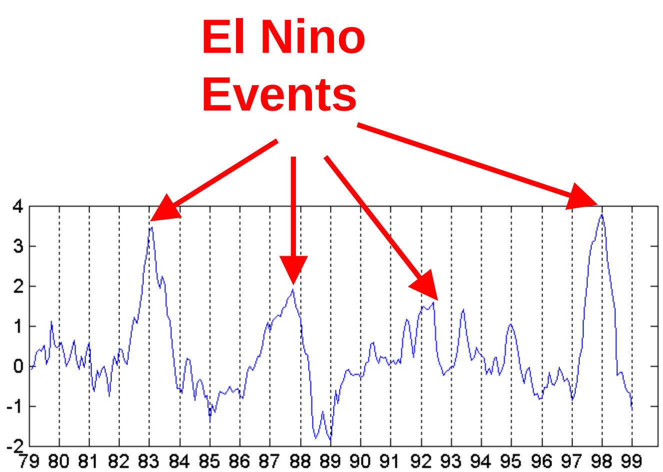
- Clustering for the detection of climate indices
- Association analysis to discover relationships between climate variables
- Automated detection of ecosystem disturbances



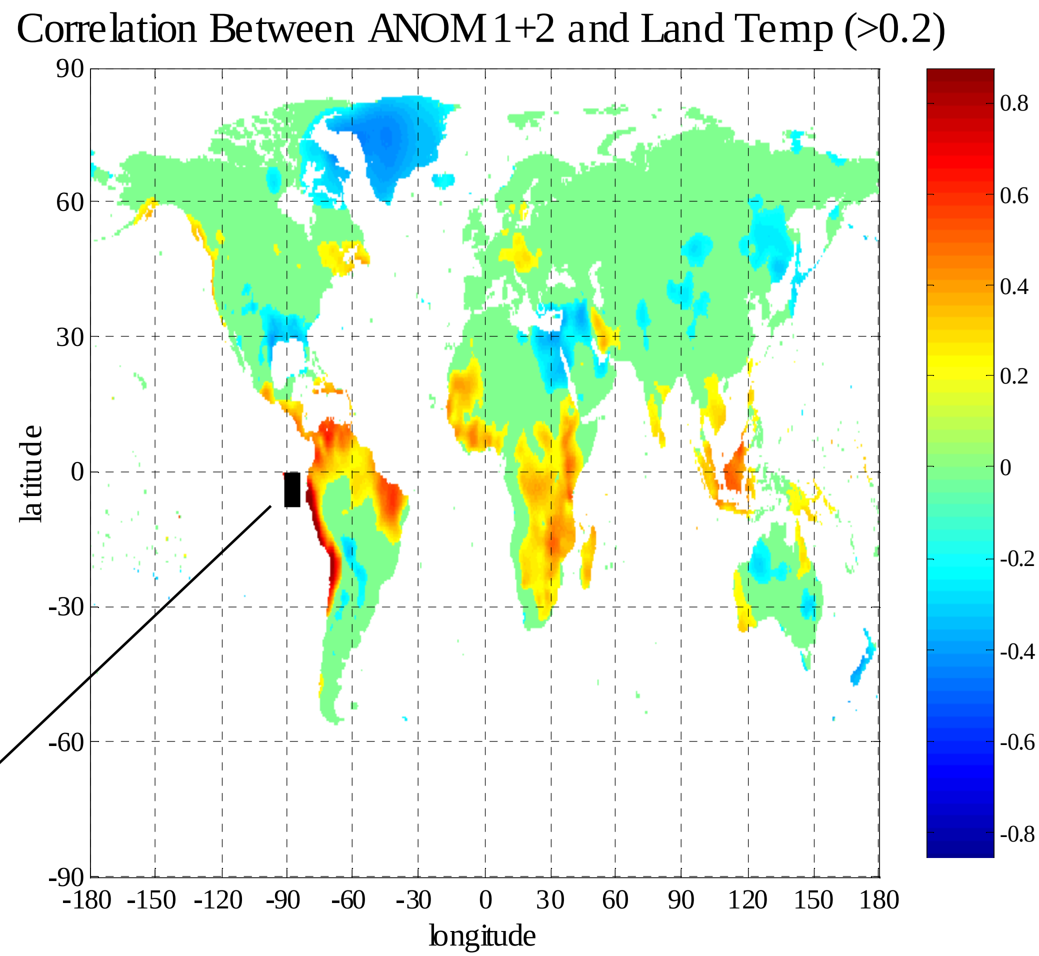
- Global snapshots of values for a number of variables on land surfaces or water
- Data sources:
  - weather observation stations
  - earth orbiting satellites (since 1981)
  - modeled-based data

Climate Indices:
Connecting the Ocean/Atmosphere and the Land

- A climate index is a time series of sea surface temperature or sea level pressure
- Climate indices capture teleconnections
  - The simultaneous variation in climate and related processes over widely separated points on the Earth

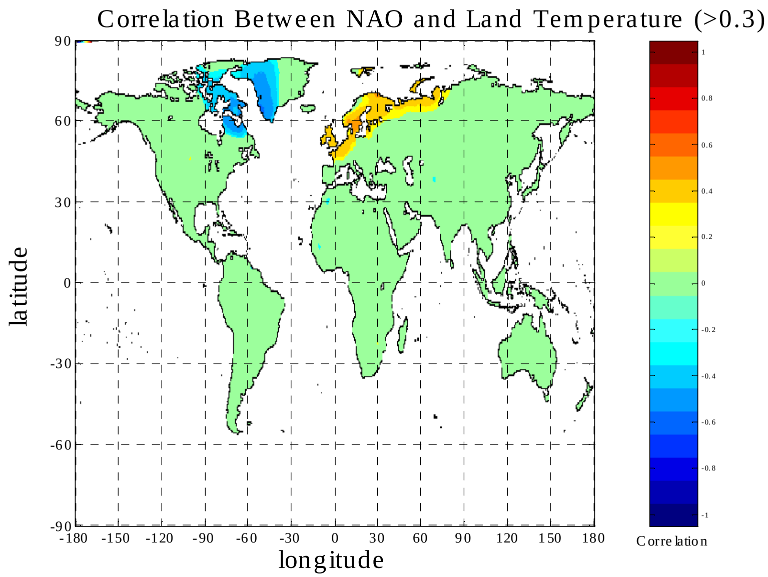


Nino 1+2 Index



Climate Indices - NAO

- The North Atlantic Oscillation (NAO) is associated with climate variation in Europe and North America.



- Normalized pressure differences between Ponta Delgada, Azores and Stykkisholmur, Iceland.
- Associated with warm and wet winters in Europe and in cold and dry winters in northern Canada and Greenland
- The eastern US experiences mild and wet winter conditions.

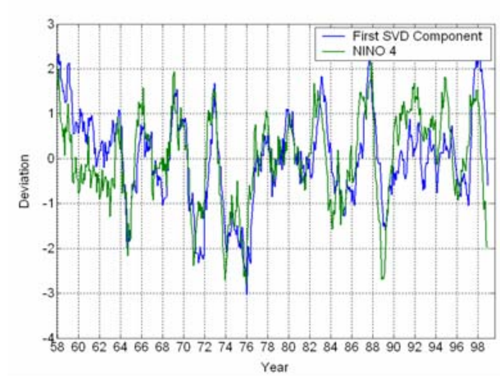
List of Well Known Climate Indices

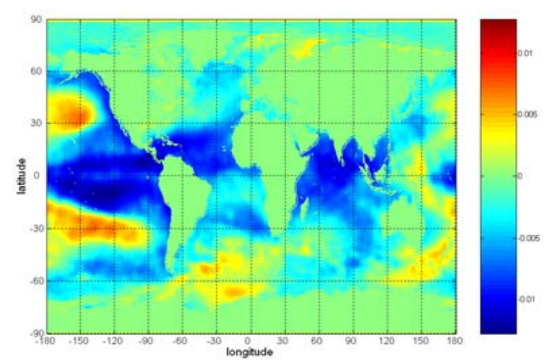
Index	Description
SOI	**Southern Oscillation Index:** Measures the SLP anomalies between Darwin and Tahiti
NAO	**North Atlantic Oscillation:** Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	**Arctic Oscillation:** Defined as the _first principal component of SLP poleward of 20° N
PDO	**Pacific Decadel Oscillation:** Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20° N
QBO	**Quasi-Biennial Oscillation Index:** Measures the regular variation of zonal (i.e. east-west) strato-spheric winds above the equator
CTI	**Cold Tongue Index:** Captures SST variations in the cold tongue region of the equatorial Pacific Ocean (6° N-6° S, 180° -90° W)
WP	**Western Pacific:** Represents a low-frequency temporal function of the ‘zonal dipole' SLP spatial pattern involving the Kamchatka Peninsula, southeastern Asia and far western tropical and subtropical North Pacific
NINO1+2	Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° -10° S
NINO3	Sea surface temperature anomalies in the region bounded by 90° W-150° W and 5° S-5° N
NINO3.4	Sea surface temperature anomalies in the region bounded by 120° W-170° W and 5° S-5° N
NINO4	Sea surface temperature anomalies in the region bounded by 150° W-160° W and 5° S-5° N

Discovering Climate Indices

- Observation
  - The El Nino phenomenon was first noticed by Peruvian fishermen centuries ago as a relationship between a persistent warm southward current around Christmas and a disastrous impact on fishing.
- Eigenvalue techniques such as Principal Components Analysis (PCA/EOF) and Singular Value Decomposition (SVD) decompose a matrix into a set of spatial patterns and a set of temporal patterns.
  - Components (patterns) must be orthogonal making physical interpretation difficult.
  - Stronger patterns tend to hide weaker patterns
  - Requires domain knowledge to select the regions of interest

We applied SVD to the global Sea Surface Temperature (SST)



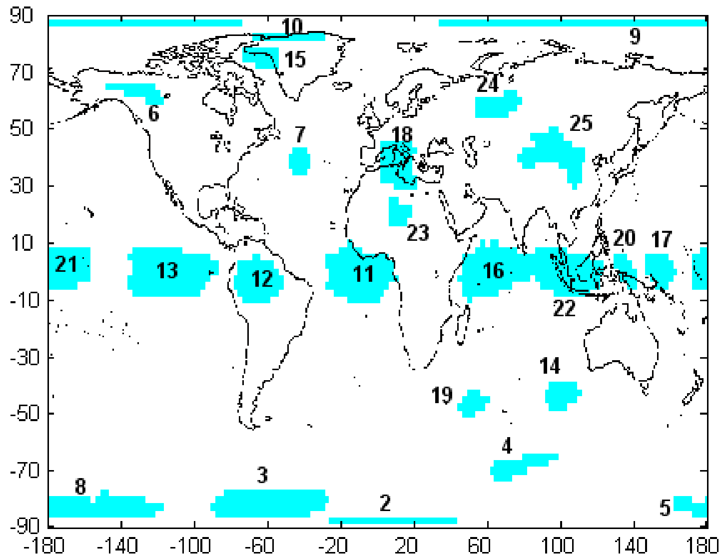


Discovering Climate Indices via Data Mining

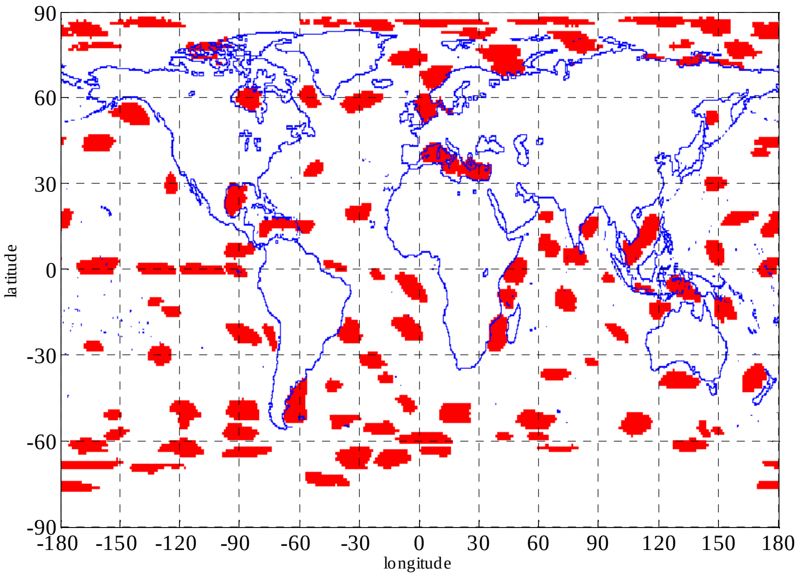
- Clustering provides an alternative approach for finding candidate indices.
  - Clusters represent ocean regions with relatively homogeneous behavior.
  - The centroids of these clusters are time series that summarize the behavior of these ocean areas, and thus, represent potential climate indices.
- Shared Nearest Neighbor clustering finds groups of points (SST or SLP time series, in this case) that have relatively homogeneous behavior.
  - Alleviates problems with varying density and problems with clusters of different shapes and sizes.
  - Can handle noisy data such as Earth Science data
  - Finds the number of clusters automatically

SLP and SST Clusters

25 SLP Clusters

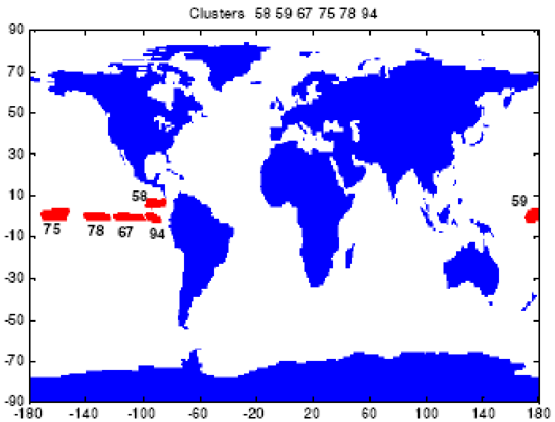


107 SST Clusters

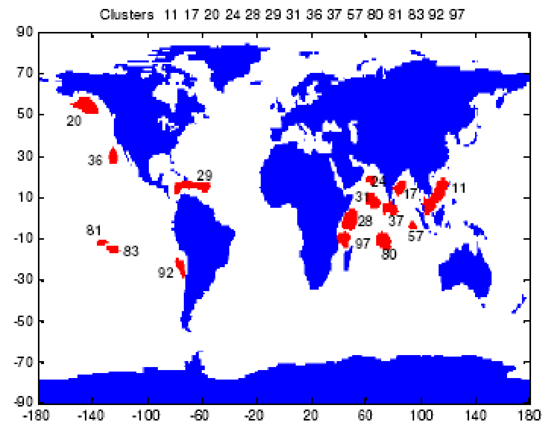


Evaluating Cluster Centroids as Potential Climate Indices

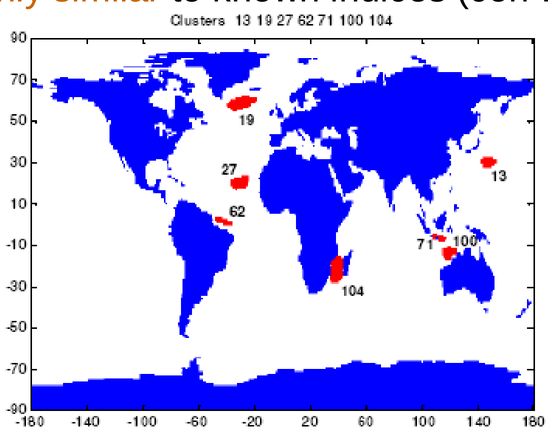
- Four cases based on similarity to known indices:



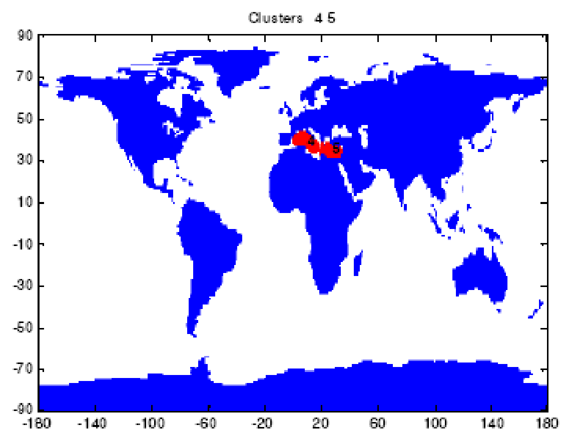
Highly similar to known indices (corr ≥ 0.8)



Similar to known indices (0.4 ≤ corr < 0.8)



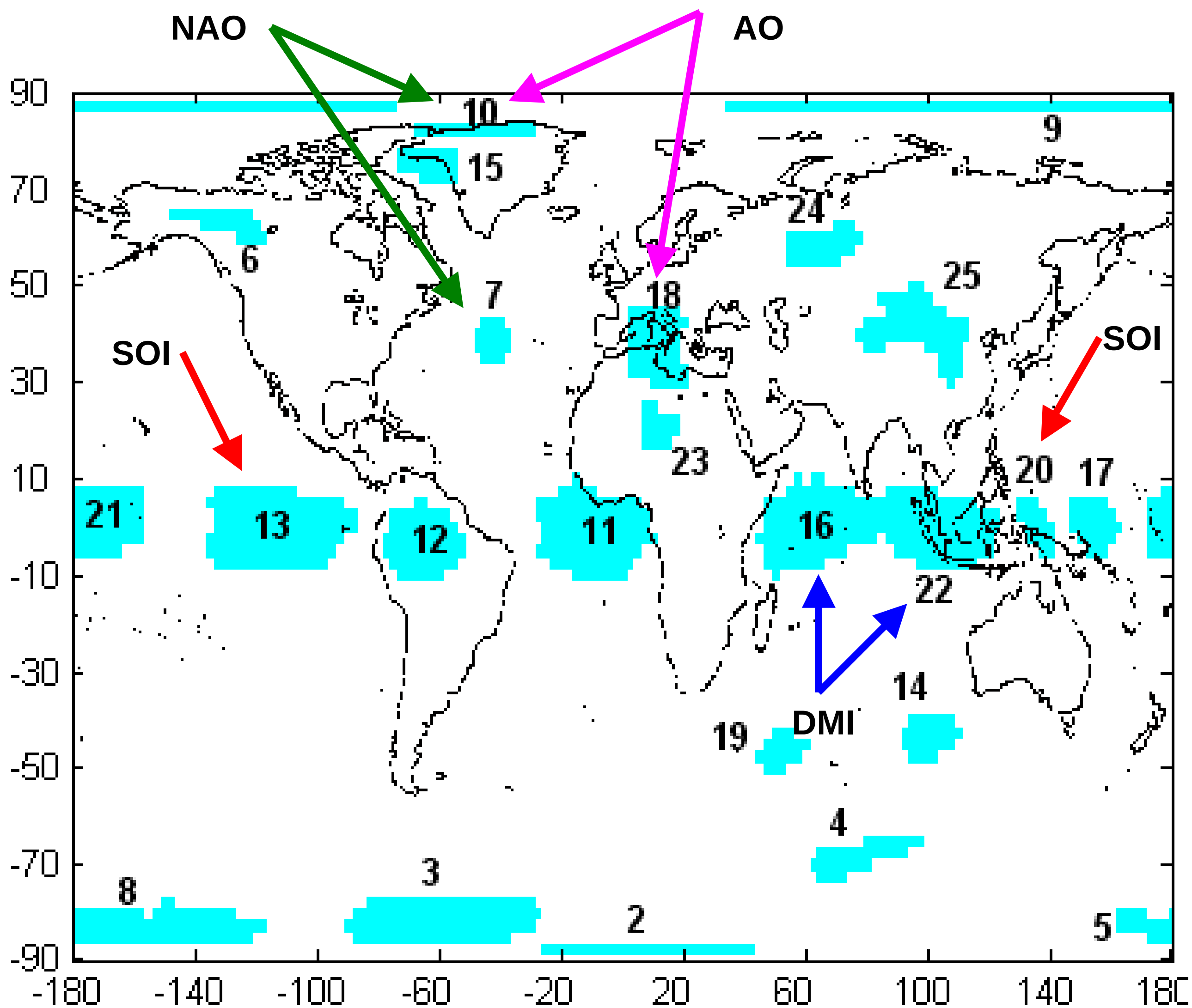
Slightly similar to known indices (0.25 ≤ corr < 0.4)



Not very similar to known indices (corr < 0.25)

An SST Cluster Moderately Correlated to Known Indices

SLP Clusters



Correlation of Known Indices with SST and SLP Cluster Centroids and SVD Components

Climate Indices	Cluster Centroids		SVD Components	
	Best-shifted Correlation	Best SST Centroid or SLP Pair	Best-shifted Correlation	Best SVD Component
SOI	-0.73	c13 - c20	0.61	3 (SLP)
NAO	0.75	c7 - c10	0.60	2 (SLP)
AO	-0.76	c10 - c18	0.82	2 (SLP)
PDO	0.52	20	-0.47	7 (SST)
QBO	-0.27	20	0.32	11 (SST)
CTI	0.91	67	-0.63	3 (SST)
WP	-0.29	c13 - c20	0.27	11 (SLP)
NINO1+2	0.92	94	-0.54	1 (SST)
NINO3	0.95	67	-0.65	1 (SST)
NINO 3.4	0.92	78	-0.68	1 (SST)
NINO 4	0.92	75	-0.69	1 (SST)

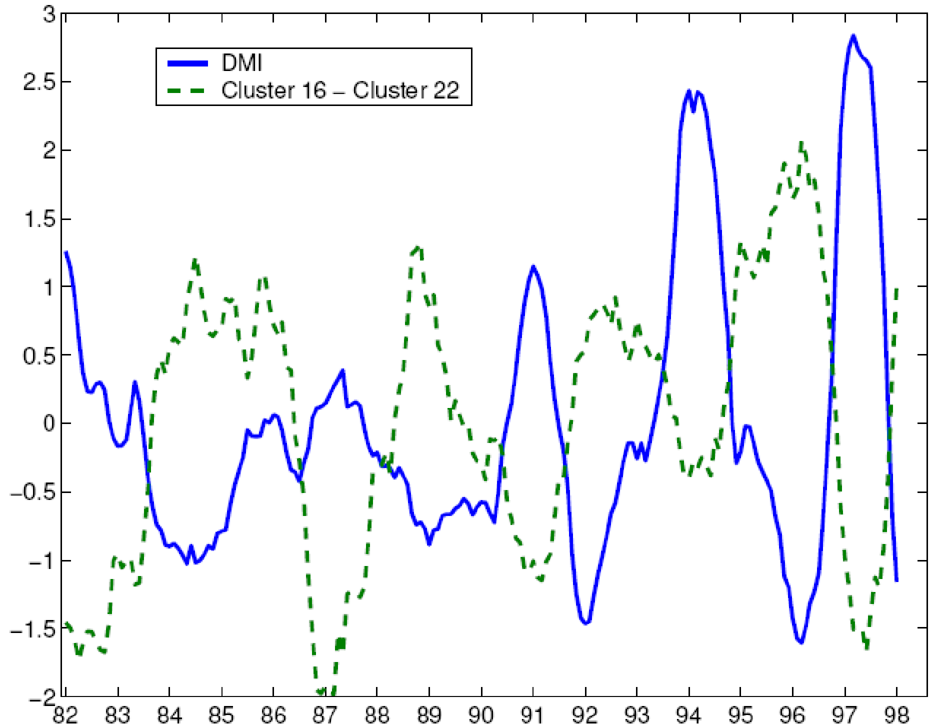
Red indicates higher magnitude of correlation.

SVD components do not have as good correlation as the cluster centroids or centroid pairs in most cases.

With some of the El Nino Indices, the leading SVD component mixes some of the indices.

Finding New Patterns: Indian Monsoon Dipole Mode Index

- Recently a new index, the Indian Ocean Dipole Mode index (DMI), has been discovered.
- DMI is defined as the difference in SST anomaly between the region 5S-5N, 55E-75E and the region 0-10S, 85E-95E.
- DMI and is an indicator of a weak monsoon over the Indian subcontinent and heavy rainfall over East Africa.
- We can reproduce this index as a difference of pressure indices of clusters 16 and 22.

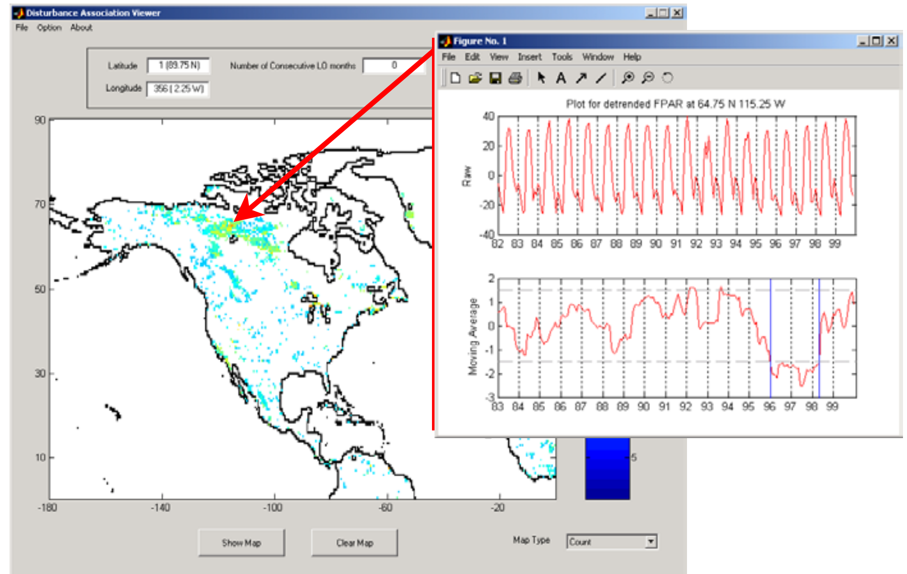


Plot of cluster 16 – cluster 22 versus the Indian Ocean Dipole Mode index. (Indices smoothed using 12 month moving average.)

Detection of Ecosystem Disturbances

Detection of sudden changes in greenness over extensive areas from these large global satellite data sets required development of automated techniques that take into account the timing, location, and magnitude of such changes.

An algorithm was designed to identify any significant and sustained declines in FPAR during an 18 year time period. This algorithm transforms a non-stationary time series to a sequence of disturbance events. Techniques were also developed to discover associations between ecosystem disturbance regimes and historical climate anomalies.



These algorithms and techniques have allowed Earth Science researchers to gain a deeper insight into the interplay among natural disasters, human activities and the rise of carbon dioxide in Earth's atmosphere during two recent decades.

NASA News

National Aeronautics & Space Administration

Ames Research Center
Moffett Field, California 94034-1000

Release: 03-51AR

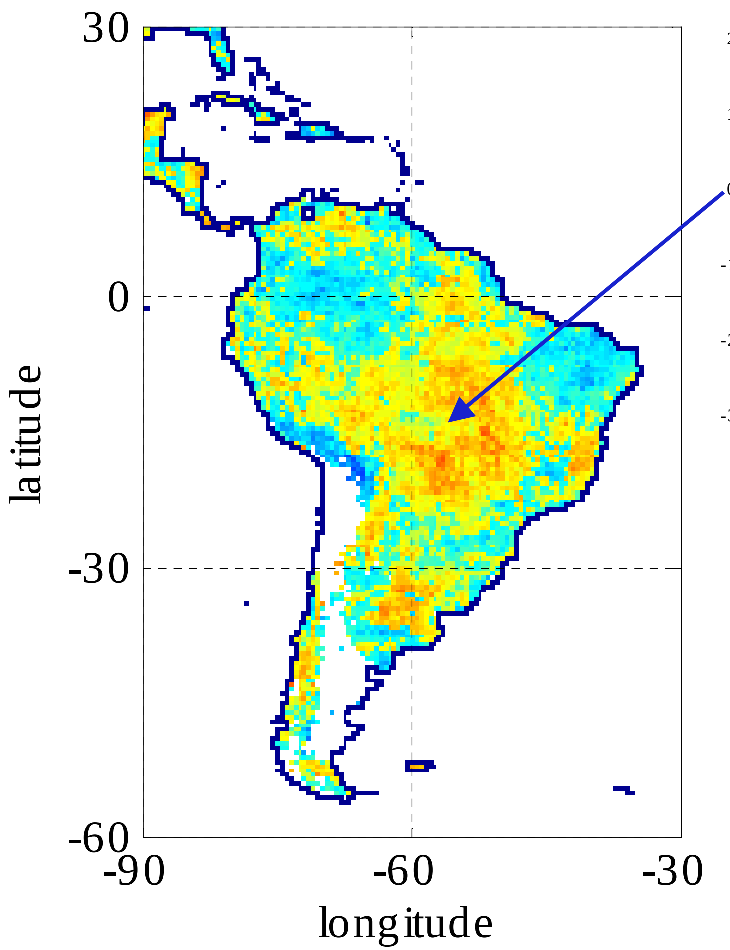
NASA DATA MINING REVEALS A NEW HISTORY OF NATURAL DISASTERS

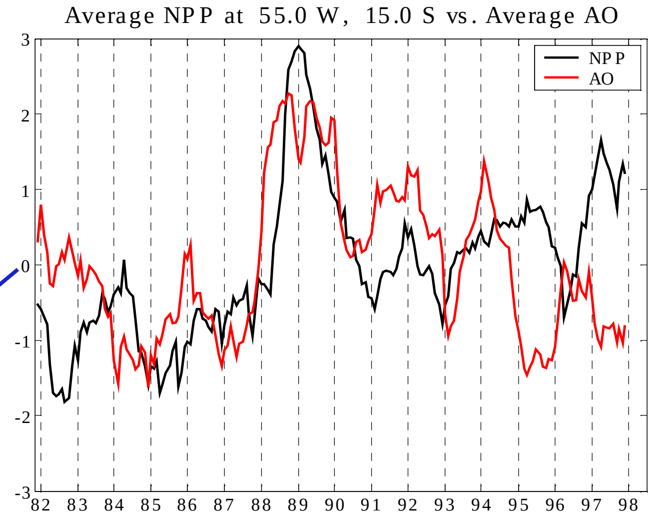
NASA is using satellite data to paint a detailed global picture of the interplay among natural disasters, human activities and the rise of carbon dioxide in the Earth's atmosphere during the past 20 years.

http://amesnews.arc.nasa.gov/releases/2003/03_51AR.html

Potter, C., Tan, P., Steinbach, M., Klooster, S., Kumar, V., Myneni, R., Genovese, V., 2003. Major disturbance events in terrestrial ecosystems detected using global satellite data sets. *Global Change Biology,* July, 2003.

Understanding Global Teleconnections of Climate to Regional Model Estimates of Amazon Ecosystem Carbon Fluxes





Potter, C. Klooster, S., Steinbach, M., Tan, P., Kumar, V., Shekhar, S. and C. Carvalho, 2002. Understanding Global Teleconnections of Climate to Regional Model Estimates of Amazon Ecosystem Carbon Fluxes. *Global Change Biology*

Discovered, using correlation analysis, a strong connection between the rainfall patterns generated by the South American monsoon system and terrestrial greenness over a large section of the southern Amazon region.

This is the first direct evidence of large-scale effects of the Atlantic Ocean rainfall systems on yearly greenness changes in the Amazon region, and the finding has important implications for the impacts of "slash and burn" deforestation on this crucial ecosystem of the world.

Conclusions

- We have demonstrated that clustering is a viable alternative to eigenvalue based approaches for the discovery of climate indices.
  - Can replicate many well-known climate indices
  - Have also discovered variants of known indices that may be "better" for some regions
  - Some indices may represent new Earth Science phenomena
  - No need for discovered indices to be orthogonal
  - No need to pre-select the area to analyze
- In our other work, we have shown that data mining can be useful for automatically detecting ecosystem disturbances, trends and associations.

Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining:

Journals

Disturbance Analysis

Christopher Potter, Pang-Ning Tan, Michael Steinbach, Steven Klooster, Vipin Kumar, Ranga Myneni, Vanessa Genovese, "*Major Disturbance Events in Terrestrial Ecosystems Detected using Global Satellite Data Sets*", *Global Change Biology*, 2003.

Teleconnections

C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, R. Nemani, and R. Myneni, "*Global Teleconnections of Ocean Climate to Terrestrial Carbon Flux*", *J. of Geophysical Research, Vol. 108, No. D17, 4556, 2003.*

C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, and C. Carvalho, "*Understanding Global Teleconnections of Climate to Regional Model Estimates of Amazon Ecosystem Carbon Fluxes*", Global Change Biology, 2003

Terrestrial Carbon Sinks

Christopher Potter , Steven Klooster, Ranga Myneni, Vanessa Genovese, Pang-Ning Tan, Vipin Kumar, *"Continental Scale Comparisons of Terrestrial Carbon Sinks", Global and Planetary Change, 39, 201-213, 2003*

C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, R. Myneni, V. Genovese, "*Variability in Terrestrial Carbon Sinks Over Two Decades: Part 1-North America*", *Earth Interactions*,2003

River Analysis

Christopher Potter, Pusheng Zhang, Steven Klooster, Vanessa Genovese, Shashi Shekhar, Vipin Kumar "*Understanding the Controls of Historical River Discharge Data on Largest River Basins* ", *Earth Interactions, 2003*

Conferences

Clustering Analysis

Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, and Steven Klooster, "*Discovery of Climate Indices using Clustering*", KDD 2003

Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, and Steven Klooster, "*Temporal Data Mining for the Discovery and Analysis of Ocean Climate Indices*", *KDD Workshop on Temporal Data Mining, 2002*

Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, and Steven Klooster, "*Data Mining for the Discovery of Ocean Climate Indices*", *The Fifth Workshop on Scientific Data Mining (2nd SIAM International Conference on Data Mining), 2002*

Vipin Kumar, Michael Steinbach, Pang-Ning Tan, Steven Klooster, Christopher Potter, Alicia Torregrosa, "*Mining Scientific Data: Discovery of Patterns in the Global Climate System* ", *Joint Statistical Meeting, 2001*

Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, Steven Klooster, Alicia Torregrosa, "*Clustering Earth Science Data: Goals, Issues and Results*", *KDD Workshop on Mining Scientific Datasets, 2001*

Association Analysis

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Christopher Potter, Steven Klooster, Alicia Torregrosa, "*Finding Spatio-Temporal Patterns in Earth Science Data*", *KDD Workshop on Temporal Data Mining*, 2001

Query Processing

Pusheng Zhang, Yan Huang, Shashi Shekhar, and Vipin Kumar, *"Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries"*, the 8th *Symp. on Spatial and Temporal Databases*, 2003

Pusheng Zhang, Yan Huang, Shashi Shekhar, and Vipin Kumar, *"Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach*", the *Seventh Pacific-Asia Knowledge Discovery and Data Mining*, 2003

Book Chapter

Pusheng Zhang, Michael Steinbach, Vipin Kumar, Shashi Shekhar, Pang-Ning Tan, Steve Klooster, and Chris Potter, Discovery of Patterns of Earth Science Data Using Data Mining, as a Chapter in *Next Generation of Data Mining Applications*, Jozef Zurada and Medo Kantardzic(eds), IEEE Press, 2003